SPEAKER INTENTION MODELING FOR LARGE VOCABULARY MANDARIN SPOKEN DIALOGUES

Yen-Ju Yang[1], Lee-Feng Chien[2], and Lin-Shan Lee[1,2]

[1]Dept. of Computer Science and Information Engineering, National Taiwan University
[2]Institute of Information Science, Academia Sinica
Taipei, Taiwan, Republic of China

Tel: 886-2-363-5251-531 Fax: 886-2-363-8247
E-mail: kathy@speech.ee.ntu.edu.tw

ABSTRACT

This paper presents a statistical speaker intention modeling approach of speech act types (SAT's)[1] prediction for large vocabulary Mandarin spoken dialogues. A SAT is an abstraction of speaker's intention in terms of the type of action that the speaker intends by the utterance. With this approach, spoken dialogue systems can be constructed to predict speaker's intention and make a proper action in advance.

1. INTRODUCTION

In spoken dialogues speaker makes conversation with computer, which realizes speaker's intention, then makes an action and generates sentences responding to the speaker's speech act types. Conventional approaches try to understand all constituents' semantic roles and relations of whole utterance by linguist's delicately designed grammar rules. But it is very difficult because of the spontaneously continuous speech recognition ambiguities and errors as well as ungrammatical sentences speaker utters. To pursue a different way of solution, we present a statistical and automatic approach to model speaker's intention by extracting features from training corpus. In addition, we design a robust word identification technique on syllable level for spotting the words in the lexicon to solve the problems of insertion, deletion and substitution from syllable recognition errors and characters inserted or deleted in a word due to extra modifiers or abbreviations. An experimental spoken dialogue system based on the proposed speaker intention modeling approach has been successfully developed and tested for a task of the telephone directory services. The test results prove that the proposed approach is efficient and can be easily applied to various spoken dialogue applications.

2. SPEAKER INTENTION MODELING

2.1. Utterance Analysis

By analyzing the speaker's utterances in a corpus obtained from a task of telephone directory services which recorded the human dialogue sessions in real-world services, we find that the constituents in the utterances can be actually classified into three major categories: the core phrases with collocated word patterns to represent the speaker's intention, the keywords in the searching database to indicate the desired data (e.g. the names of the subscribers in this telephone directory services task), and the redundancy words which are irrelated to the speaker's intention. For example, four typical inquiries in the corpus are illustrated below:

- Please help me find out that phone no. of (Taiwan Bank), thanks!
  請幫我查那個（台銀）的電話號碼，謝謝！
- May I ask the no. of (City Bank)?
  可不可以請問（花旗銀行）的號碼？
- Are there other numbers?
  有其它號碼嗎？
- Please give me another line.
  請給我別線。

In these examples, the underlined words are core phrases, the parenthesized words are keywords, and the remaining words are redundancy. In fact, the first two examples have a similar intention to inquire the telephone numbers of a certain subscriber in this task. And, the last two examples are the same to search for another telephone number. For further analysis, the inquiries with a similar intention are classified into the same SAT. This will be useful in modeling speaker's intentions by a statistical approach.

2.2. Core Phrases Extraction

Since we find that the utterances with the same SAT are constituted of certain core phrases and types of keywords, each

SAT then can be modeled by its composed core phrases and types of keywords regardless of the interference of the redundancy words from a large vocabulary. The core phrases and keywords of a SAT make up the lexicon of the SAT in our experiments. As the keywords are easier to be identified in the database, the extraction of the core phrases are more important. With our approach, the core phrases of each SAT are extracted automatically from the part of the corpus, which has been manually labeled the SAT's. The initial set of the acquired core phrases is composed of the words with high frequencies. Then we extract the collocation word pairs which are measured by mutual information shown in formula (1).

$$I(x,y) = \log \frac{f(x,y)}{f(x)f(y)} \tag{1}$$

When any concatenated patterns (x, y) has high mutual information and $f(x)$ as well as $f(y)$ are greater than a constant c in order to avoid overestimation, (x, y) is added into the set, and the original patterns, x and/or y, will be deleted if existing in the set. Viewing theses word pairs as new words, we can extract the collocation word triplets by the same way. This process is repeated executed to extend the length of collocation word patterns until convergence, i.e., no longer patterns extracted [2].

2.3. Automatic SAT Annotation

Furthermore, we try to annotate the SAT for each sentence in the other large part of corpus for training a language model. Based on the maximum likelihood estimation, the SAT of a sentence S, SAT_S^*, is determined by formula (2),

$$\begin{aligned} SAT_S^* &\stackrel{def}{=} \arg\max_{SAT} score(SAT|S) \\ &\stackrel{def}{=} \arg\max_{SAT} \max_{\text{path } W} \sum_{\substack{w_i \in Lexicon(SAT) \\ w_i : occur(w_i) \geq 1/2 \\ w_i \in W}} length(w_i) \end{aligned} \tag{2}$$

that is, for each sentence in the test corpus only the SAT with maximum score is labeled. The likelihood score here is the sum of the length of w_i in the longest word path $W = \{w_1, w_2, \ldots\ldots, w_n\}$. Because of the existence of extra modifiers and abbreviation, conventional exact string matching techniques for word identification can't work well here. For example, there is a word (core phrase) in the lexicon such as "find that", but there also have similar phrases, such as "find out that", "find out" or "find", not extracted in lexicon but appearing in S. For this reason we define the estimation of occur rate in formula (3).

$$occurr(w) = \frac{\text{number of syllables of } w \text{ occurred in } S}{\text{number of syllables of } w} \tag{3}$$

Several examples of occur rate are shown in Figure 1, where we use different graphs to represent different syllables. If the occur rate is above or equal to 1/2, this word can be spotted. By this way, though some of the spotted words may overlap, connecting nonoverlapping words forms some paths. The length of longest word path determines the score of each SAT. Using formula (2) to annotate SAT of the corpus, the accuracy of almost 100% can be achieved in our experiments.

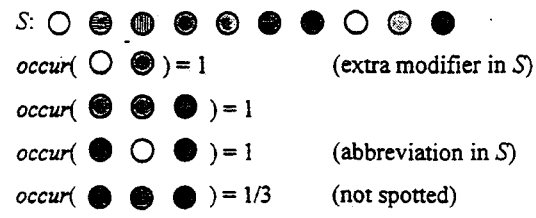

Figure 1: Examples of the estimation of occur rate

2.4. Language Modeling

Because of the speech recognition ambiguities and errors, many wrong word hypotheses may be spotted. The association of many adjacent word hypotheses may be very weak, we can not strongly believe the longest path in word lattice is the optimal path. It is well known that the word N-gram language model used in the linguistic processing of speech recognition achieves good performance. We therefore try to define an appropriate semantic language model for SAT prediction of continuous speech. The keywords, representing inquired data belonging to some fields of database, can be viewed as conceptual categories (CCs) depending on the fields. The core phrases indicating the speaker's intention also can be classified into some CCs with similar usage. Then we can train a CC-based language model for every SAT used in prediction. So as to, we want to automatically cluster the core phrases into some CCs. First, we create the multi-dimensional feature vector for each of the core phrases. Each dimension means the count of coocurrence with different successor and the number of dimensions is all core phrases plus the number of CCs of keywords. Second, using a modified k-means clustering algorithm [3] to classify the core phrases into CCs. The CC bigram Markov language model, whose parameters are trained from the SAT annotated corpus with redundancies deleted, is defined in formula (4).

$$P(W) = \prod_{i=1}^{n} P(w_i|C(w_i))P(C(w_i)|C(w_{i-1})) \tag{4}$$

With this CC bigram Markov language model, we can apply it to not only the SAT prediction from continuous speech described in the following section but also annotating SAT of text corpus, that is, the $score(SAT|S)$ in formula (2) can be defined as the maximum likelihood Markov chain instead of the longest path.

path. We believe the annotation can be more reliable.

3. SAT PREDICTION OF SPEECH

For an arbitrary input speech utterance, by utilizing the mono-syllabic structure of Chinese language, the syllable recognizer [4] is adopted to first provides a set of syllable candidates for each input syllable to construct a syllable lattice, on which the "words" in the "lexicon" of each SAT are spotted and the maximum likelihood path is found by an integration with the CC bigram Markov language model. The estimated top n SAT's are then further processed by a dialogue manager.

In order to solve the problems of insertion/deletion/ substitution from syllable recognition errors and extra modifiers or abbreviations, we design a robust word identification technique on syllable level for spotting words. For each word, we generates a "position index graph" recording the positions each syllable occurred on syllable lattice, then align word boundary by searching the shortest path on graph with a Viterbi algorithm. Two examples are shown in Figure 2. The cost of each node in graph is determined by "trigram distance", i.e. an accumulated distance with preceding two nodes. For implementation of such a trigram distance, the Viterbi search is performed on a finite state network (FSN) as shown in Figure 3, and the estimation of the cost is defined in formula (5),

$$\begin{aligned} &cost(state_{i,m,n}) \\ &= \min_{l}\left[cost(state_{i-1,l,m}) + d(p_{i,n}, p_{i-1,m}, 1) + d(p_{i,n}, p_{i-2,l}, 2)\right] \quad (5) \\ &d(a,b,c) = abs(a-b-c) \end{aligned}$$

where $p_{i,n}$ is the position index of ith syllable's nth candidate in position index graph. Those words with cost lower than the threshold are selected as spotted hypotheses. We further define the possibility score of words by integration of the costs and acoustic scores as shown in formula (6),

$$WordScore = w_1 * cost * word\ length + w_2 * AcousticScore \quad (6)$$

where $w_1 < 0$, because the cost is a minimum estimation. Connecting nonoverlapping word hypotheses forms a word lattice, where the maximum likelihood path is found by integration of the CC bigram Markov language model and *WordScores*. So, the possibility score of each SAT for an input speech utterance U, is defined in formula (7),

$$\begin{aligned} &score(SAT|U) \\ &\overset{def}{=} \max_{W:\ \substack{W=(w_1,\ldots,w_n) \\ w_i \in Lexicon(SAT)}} \left[\lambda_1 \log P(W) + \lambda_2 \sum_{w_i} WordScore(w_i)\right] \end{aligned} \quad (7)$$

where $P(W)$ and $WordScore(w_i)$ are defined in formula (4) and (6) respectively. The top n SAT's are then further processed by the dialogue manager.

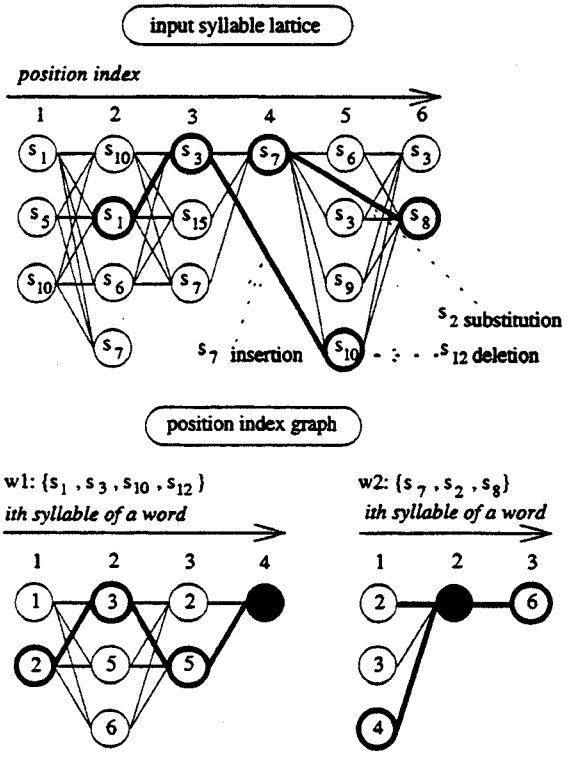

Figure 2: The examples of position index graph

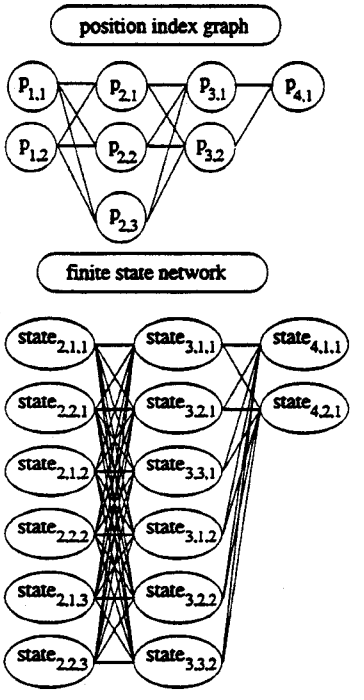

Figure 3: The FSN converted from position index graph

4. EXPERIMENTAL RESULTS

An experimental system based on the above approach has been successfully developed for a task of telephone directory services for banking/financing organizations in Taipei. There are about 500 dialogue sessions with about 4000 utterances (roughly half by users and half by the operators) taken as training corpus. After analyzing the corpus, the possible SAT's are listed in Table 1. Besides, two speakers are selected to test in the preliminary experiments. The obtained accuracy of SAT prediction are listed in Table 2 and Table3 respectively, where it can be found that although the syllable inclusion rates of speaker B is lower, the achieved accuracy is close to that of A.

1	inquire: a phone no.
2	answer: the desired company
3	answer: the desired dept. or address
4	echo: 7~10 digits
5	request: repeating the no. or express unclear hearing
6	request: total no.
7	request: other no.
8	doubt: the no. is not working
9	request: extension no.
10	express: some beginning digits known
11	answer: confirming system response or yes/no question
12	echo: repeating system response
13	greetings
14	ask for help
15	thanks
16	say good-bye

Table 1: The speaker's SAT

no. of syll. included	1	5	10	15	20
syll. incl. rate	75.70%	82.86%	83.38%	83.38%	84.14%
top 1 SAT accu.	87.80%	85.37%	80.49%	80.49%	80.49%
top 2 SAT accu.	95.12%	95.12%	95.12%	95.12%	95.12%
top 3 SAT accu.	100%	100%	100%	100%	100%

Table 2: The SAT accuracy for speaker A

no. of syll. included	1	5	10	15	20
syll. incl. rate	63.17%	80.31%	82.86%	83.38%	83.89%
top 1 SAT accu.	70.73%	82.93%	78.05%	78.05%	80.49%
top 2 SAT accu.	75.61%	92.68%	92.68%	92.68%	95.12%
top 5 SAT accu.	90.24%	100%	100%	100%	100%

Table 3: The SAT accuracy for speaker B

5. CONCLUDING REMARKS

The proposed approach is capable of modeling speaker's intention by extracting features from training corpus instead of using delicately designed grammar rules. The test results prove that the proposed approach is efficient and can be easily applied to various spoken dialogue applications. In addition, the used robust word identification technique is able to reduce the problems of insertion/deletion/substitution from syllable recognition errors and extra modifiers or abbreviations.

ACKNOELEDGMENT

The authors acknowledge the support of Telecommunication Laboratory, MOCT, Taiwan, R. O. C., under contract TL-85-5205. We also thank useful discussions with E. F. Huang and C. J. Lee.

6. REFERENCES

[1] M. Nagata and T. Morimoto, "An Experimental Statistical Dialogue Model to Predict SAT of the Next Utterance," Proc. of *ISSD*, pp. 83-86, 1993.

[2] F. Smadja, "Retrieving Collocation From Text: Xtract," *Computational Linguistics*, Vol. 19, No. 1, pp. 143-178, 1993.

[3] G. Jay, "A Modified K-Means Clustering Algorithm for Use in Isolated Word Recognition," IEEE Trans. on ASSP, Vol. ASSP-33, NO. 3, June, 1985

[4] H. M. Wang, Y. J. Yang, L. S. Lee, *et. al.*, "Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary but Limited Training Data," Proc. of *IEEE ICASSP*, pp. 61-64, 1995.